McKinsey & Company

Case Study

An AI power play: Fueling the next wave of innovation in the energy sector

How Vistra Corp. is partnering with McKinsey to improve efficiency and reduce emissions by using AI

May 2022

Tatum, Texas might not seem like the most obvious place for a revolution in artificial intelligence (AI), but in October of 2020, that's exactly what happened. That was when Wayne Brown, the operations manager at the Vistra-owned Martin Lake Power Plant, built and deployed a heat rate optimizer (HRO).

The "heat rate" is basically the amount of electricity generated for each unit of fuel consumed. To reach the optimal heat rate, plant operators continuously monitor and tune hundreds of variables, or "set points" on things like steam temperatures, pressures, oxygen levels, and fan speeds. It's a lot for any operator to get right 100 percent of the time—so Vistra thought AI could help.

With this goal in mind, Wayne and his group worked together with a McKinsey team that included data scientists and machine learning engineers from QuantumBlack[1] AI by McKinsey, to build a multilayered neural-network model—essentially an algorithm powered by AI that learns about the effects of complex nonlinear relationships. This model went through two years' worth of data at the plant and learned which combination of external factors—such as temperature and humidity—and internal decisions, like set points that operators control, would optimize the algorithm and attain the best heat-rate efficiency at any point in time.

Vistra team members provided continuous guidance about the intricacies of how the plant worked, and identified critical data sources from sensors, which helped McKinsey engineers refine the model, adding and removing variables to see how those changes affected the heat rate.

Through this training process, and by introducing better data, the models "learned" to make ever more accurate predictions. When the models were accurate to 99 percent or higher and run through a rigorous set of real-world tests, a McKinsey team of machine learning engineers converted them into an AI-powered engine. This generated recommendations every 30 minutes for operators to improve the plant's heat-rate efficiency. At a meeting with all of Vistra's leaders to review the HRO, Lloyd Hughes, a seasoned operations manager at the company's Odessa plant, said, "There are things that took me 20 years to learn about these power plants. This model learned them in an afternoon."

With this kind of power at their fingertips, Wayne and his team could make better, more informed decisions. Acting on the HRO recommendations helped Martin Lake run more than two percent more efficiently after just three months in operation, resulting in $4.5 million per year in savings and 340,000 tons of carbon abated. This carbon reduction was the equivalent of taking 66,000 cars off the road.[2] If that doesn't sound like a lot, consider this: companies that build gas-fueled power plants invest millions of dollars in research and development over four to five years to achieve a one-percent improvement in power-generation efficiency. Vistra hit that improvement level in only one-twentieth the amount of time using the data and equipment it already had.[3]

Vistra has since rolled the HRO out to another 67 power-generation units across 26 plants, for an average one-percent improvement in efficiency, and more than $23 million in savings. Along with the other AI initiatives, these efforts have helped Vistra abate about 1.6 million tons of carbon per year, which is ten percent of its remaining 2030 carbon-reduction commitment. That's equivalent to offsetting about 50 percent of what a 500-megawatt coal plant emits.

What happened at Martin Lake has happened at dozens of Vistra's other power plants, with more than 400 AI models (and counting) deployed across the company's fleet to help operators make

[1] QuantumBlack is a McKinsey company.
[2] Calculations based on "Greenhouse Gases Equivalencies Calculator - Calculations and References", United States Environmental Protection Agency
[3] Calculations based on source material from "A Brief History of GE Gas Turbines", Power, July 8, 2019, and Frequently asked questions, US Energy Information Administration, last reviewed September 20, 2021
[4] The tree equivalency is based on calculations derived from the "The global tree restoration potential" study, published in the journal Science, July 5, 2019.

About Vistra and its emissions-reduction goals

Vistra Corp. is the largest competitive power producer in the United States and operates power plants in 12 states with a capacity of more than 39,000 megawatts of electricity—enough to power nearly 20 million homes.

Vistra has committed to reducing emissions by 60 percent by 2030 (against a 2010 baseline) and achieving net-zero emissions by 2050. To achieve its goals, the business is increasing efficiency in all its power plants and transforming its generation fleet by retiring coal plants and investing in solar- and battery-energy storage, which includes the world's largest grid-scale battery energy-storage facility. Vistra's to path to net zero will also require it to grow its zero-carbon portfolio to more than 7000 megawatts by 2026.

even better decisions. It also reflects a core trait of Vistra's AI transformation, which is that it isn't a story of one massive hit, but rather the story of dozens of meaningful improvements snowballing to deliver significant value in terms of accelerating sustainable and inclusive growth. It's also the story of how an organization architected an approach to rapidly scale every successful AI solution across the entire business. And it's a story of how a continuous improvement culture, combined with a powerful AI modeling capability, helped leaders and plant operators do their jobs better than ever before.

With more than $60 million captured in about one year of work and another $40 million in progress, Vistra is well on its way to delivering against a roadmap of $250-$300 million in identified EBITDA and more than two million tons of carbon abatement per year. The AI-driven advances at Vistra have heralded a generational shift in the power sector in terms of improvements in efficiency, reliability, safety, and sustainability.

If the one-percent improvement in efficiency the HRO delivered across the fleet was carried across all coal- and gas-fired plants in the US electric-power generation industry, 15 million tons of carbon would be abated annually—the equivalent of decommissioning more than two large coal plants or planting about 37 million trees.[4] That means less fuel needed to deliver power to the hospitals, schools, and businesses that rely on it. AI has the potential to bring similar levels of improvement to renewables as well, making them a more cost-effective and attractive energy option.

Turning points on the AI journey

Healthy skepticism and a culture of favoring action over words at Vistra meant that the biggest hurdle in the AI journey wasn't the technology: it was the people. Vistra leadership and operations managers needed to know what AI could do and be convinced it could really work.

It was this ingrained culture of continuous improvement—alongside a highly competitive market and a commitment to sustainability—that convinced Vistra's leadership they needed to give AI a chance.

What does "machine learning operations" mean?

Machine learning operations is the set of practices and infrastructure to manage the production and deployment of AI solutions or products. Improvements in AI tooling and technologies have dramatically transformed AI workflows, expedited the AI application life cycle, and enabled consistent and reliable scaling of AI across business domains. This framework enables organizations to create a standard, company-wide AI "factory" capable of achieving scale.

Seeing real possibilities

The first question was a relatively simple one: "How can AI help improve the way Vistra generates power?"

Answers to that question bubbled over when 50 of Vistra's top leaders came to a McKinsey-hosted workshop. In multiple sessions, experts explained how AI worked, walked through detailed case studies showing how other companies used analytics and AI to generate value and gave live demonstrations of technologies, including digitized workflows and machine learning. Leaders in analytics from various sectors—including Amazon, Falkonry, Element Analytics, and QuantumBlack, AI by McKinsey, as well as G2VP from the venture-capital world—provided insights and examples of how AI works.

"I saw an example of how a metallurgical plant was using AI to help its operators optimize set points and it clicked for me," remembers Patrick "Cade" Hay, the plant manager for Vistra's Lamar power plant. "I saw how I could translate that into helping me run my own plants more efficiently. This was my lightbulb moment."

Company leaders and plant managers poured over process flow sheets and engineering diagrams to determine pain points as well as opportunities. This exercise allowed them to focus first on finding where the value was, then secondly on what technologies were needed to deliver it. Many of the operations opportunities were around yield and energy optimization and predictive maintenance, which according to our research, were the top AI use cases for manufacturing industries.[5]

By the end of the session, Vistra had developed a strategy to develop a series of AI solutions that could capture $250-$300 million in potential EBITDA while helping the company achieve its 2030 emissions-reductions goals.

Codeveloping the AI

While the analysis looked promising, proving it in the field was what mattered. "If our plant managers aren't bought in, then things don't happen," explained Barry Boswell, Vistra's executive vice president of power-generation operations. "So, we

[5] "The state of AI in 2021", a McKinsey Global Survey, December 8, 2021.

said, 'Let's pick a leader who is knowledgeable and skeptical, because if we can win them over, we can get everyone.'"

They picked Cade. Not only did he run a top-performing plant in terms of profitability, reliability, and heat rate, he had a reputation for telling it like it is—Barry trusted Cade to tell him whether or not the value potential in AI was real. When he approached Cade about testing out a proof of concept to optimize duct burners, he was intrigued but predictably skeptical. Cade saw the potential in AI but was interested in finding out if it could actually help in the field.

Duct burners essentially work like afterburners in jet planes; they provide a surge of energy when needed. Operators use them as supplements to hit energy targets, which are known as their "dispatch point."

The issue is that powering duct burners uses more fuel than regular methods, so it's more expensive, generates more carbon emissions, and increases the wear and tear on equipment.

McKinsey subject matter experts, data scientists, and analytics translators from QuantumBlack, AI by McKinsey, worked closely with a team from Vistra comprised of power generation and process experts as well as front-line operators to understand how the plant works, what data was available from the sensors already in place, and what variables could be directed—like the fact that the number of cooling fans running could be controlled, while the ambient temperature couldn't.

As the teams developed the models, plant operators reviewed recommendations to see what made sense, what other variables needed to be tested, and what kinds of recommendations the operators

AI solution deep dive

QuantumBlack's Ayush Talwar, an expert associate partner, provides more details on the AI model development:

What kind of models did you use?
We used a range of models to fit the specific needs of the plant and the solution. These ranged from Bayesian regression models to deep-learning models.

How did you choose which model to use?
We wanted to find the right balance between model performance, meaning its accuracy; explainability, which describes relationships to the operators; the source richness, which refers to the amount of data available; potential actionability, or how many of the most important features in the model people can actually control; and maintainability, which will reduce need for more technical skills to maintain them.

How would you compare the models you used to older ones?
Our models were more accurate than previous models, and importantly, more actionable. We used a variety of metrics to measure model performance such as the mean absolute percentage error (MAPE) and root mean squared error (RMSE). Our models typically achieved MAPEs of less than 1.5 percent.

How did you build up your models?
Each neural network model was made up of several layers with each layer containing batch normalization, dropout, and activation. We created cross validation and out-of-sample testing sets using time-based splits to prevent overfitting models. We added features based primarily on domain expertise.

How did you turn models into tools operators could use?
We used machine learning models to make accurate predictions of outcomes. We then wrapped meta-heuristic optimization algorithms around those models to generate recommendations that helped operators make decisions about how to run the plant better. The models were then embedded within the existing production workflows and deployed live in the operator room.

Working together

Denese Ray, operations shift supervisor for Vistra's Coleto Creek power plant; Doug Richter, the shift supervisor at the same plant; and Muro Kaku, McKinsey analyst and data scientist, discuss how they worked together.

Denese: "Every week, I'd bring in a different supervisor to join us for our check-in calls with McKinsey. In this way, Muro and the McKinsey team saw what we were dealing with. We really needed all our supervisors and their feedback to build the model so we could get the most from it."

Muro: "That really mattered. For example, when we first trialed the HRO at the Coleto Creek plant, the recommendations were frequently rejected by the operators. We didn't know why. When we spoke with the operators, they told us about specific rejection reasons, such as the tool recommended increasing superheat temperature by 30 degrees in a 15-minute interval to maximize the heat rate, but it takes the plant longer to heat up, so they rejected the recommendation. We then added a constraint to the tool to keep the increase and decrease to five degrees in a 15-minute interval, which operators could achieve."

Doug: "During load changes [increase or decrease in power output], the parameters are changing constantly, so recommendations to increase air flow at that time were counterproductive. We worked with McKinsey to have the recommendations stop for 30 minutes while we were changing load until the plant was stable again."

Denese: "We were all pretty skeptical of the tool at first, but when we got to play with the heat rate optimizer and see how it worked, and how well McKinsey worked with us in the plant, we understood how it could help."

would find most helpful. By analyzing the effect of various inputs and set points on the plant—such as pressure and humidity, the angle of blades in the gas turbine, usage of inlet cooling technologies, and the age and performance of various components like filters and condensers—and running it through the model, the analysis was clear: overall duct-burner usage could be reduced by approximately 30 percent, which would result in the equivalent of $175,000 of yearly savings on fuel costs and wear and tear on the system, in addition to an abatement of about 4,700 tons of carbon per year.

"We worked closely with the team from McKinsey to develop AI models that reflected the realities of how power plants operate," said Cade, "and then when

Going for scale and adoption from the beginning

Vistra's leadership realized from the beginning that the only way to achieve their efficiency and carbon-abatement aspirations was to scale every solution. "We manage the Vistra fleet as one. If a plant is doing something that works, we want every plant to do it," says Barry. "That's what we're built to do."

That realization led Vistra to invest in a five-part system to scale and sustain AI solutions:

1. Turn each successful proof of concept (or MVP) into a product

A standardized solution would need to be deployed at each power plant and would become easier to maintain over time. When a solution has proven value at a pilot site and is approved for scaling, a team of software and machine learning engineers immediately takes over to refactor, modularize, and containerize the code. That way, there is a single software "core" package for each deployment that can be updated and improved. A product owner manages the overall process and takes ownership for use and adoption.

Over time, the team developed seven solution archetypes, which provided consistent approaches, logic, assumptions, and algorithmic elements as a basis for each new application being developed. This gave each new solution a big head start when development began. It took ten-to-12 weeks to build the first HRO. Rolling each HRO out to subsequent new plants now takes just two-to-three weeks.

2.Create machine learning operations infrastructure

Vistra implemented a machine learning approach to essentially create a "factory" that standardized the deployment and maintenance of more than 400 AI models. At a high level, this approach enabled the team to bring live data from each of Vistra's power units into a single database, use GitLab software to manage version control for code, containerize the code so it could be easily deployed to any environment, set up a scheduler—using Apache Airflow—to make sure recommendations and actions were delivered on time, create dashboards to monitor model performance and usage, and manage the continuous improvement of each model to make sure the plants were sustaining captured value.

Teams also incorporated multiple approaches to reduce risk by building functional limits, such as maximum throttle pressure or heat levels into the code, and put all code through biweekly peer reviews and multiweek testing. McKinsey's risk-dynamics experts worked with the team to test assumptions, review code, and ensure that all risk best practices were reflected in the models.

3.Customize and adapt for "last-mile" implementation

Because each product is designed to be modular and reusable, 50-70 percent of each one is ready from day one to be used when it's rolled out to a power plant. But customization is always needed because each plant has its own unique characteristics.

Take the maintenance solution developed to understand the best time to replace inlet filters at gas turbines. Each filter costs $150,000, and a plant needs to be shut down for two days to replace one. At the Moss Landing Power Plant in California, the unit that faces the ocean had to deal with a lot more moisture and salt than the one facing inland, so the degradation profile for each filter was different. Similarly, plants in Texas have to manage for dust, while those in Connecticut battle cold weather—local conditions create their own unique degradation profiles.

So, dedicated customization teams made up of data scientists, engineers, operators, and power-generation experts worked with each power plant to tailor the solution to the unique conditions of that particular plant.

4. Build capabilities

Building on Vistra's well-established culture of continuous improvement, McKinsey worked to help Vistra get the most from AI. That included running a two- to three-week program for operators at each plant to explain the models and how they were developed, as well as instructions on how to use the tool itself.

As Rachit Gupta, Vistra's vice president of generation and wholesale technology applications said, "People had to know what the model was doing and learn to trust that it was right. Once they saw that the models were generating recommendations that made sense and lowered heat rate, they were ready to start using them more."

Additional training for the core tech team covered how to build and maintain models and resolve issues, as well as deeper machine learning and analytics skills to understand how models work in detail. This happened largely by working side by side with McKinsey AI experts.

5.Design for the operator experience

From the beginning of the solution-development process, McKinsey designers worked with operators to understand what their day-to-day activities look like. What soon became clear was that plant operators had real constraints on their time and had to manage dozens of inputs tracked on an array of screens in the control room. Adding to that workload would be a sure-fire way to overwhelm operators and reduce the effectiveness of the solutions. The tools had to make operators' lives easier.

For this reason, the screen that displayed the AI solutions and recommendations were integrated into PowerSuite, an interface that operators already used, so they didn't need to monitor yet another screen. The displays themselves were designed to be easy to read. A solution displays a green signal if the plant is running optimally for the given conditions, and red when it isn't. A brief recommended action accompanies any red display, with the value attached for implementing that recommendation.

"There's a simple screen that not only shows you what needs to be optimized, but how much it's worth," says Lloyd. "That really registers with people."

Vistra's story is far from finished. As Cade put it, "We're just looking at the tip of the iceberg." Vistra's roadmap for 2022 and 2023 includes bringing AI to its rapidly growing renewables fleet of solar and batteries to optimize yield and reliability, among other initiatives.

To help sustain this ambition, Vistra is building up its talent bench. In addition to hiring a small team of data scientists and engineers, Rachit has partnered with the University of Texas at Dallas to offer basic, intermediate, and advanced courses in AI and analytics for Vistra employees. Some 70 people have already completed courses, including those reskilling from statistics to machine learning. Vistra has also built relationships with local colleges and universities to develop internship programs and work with students in capstone projects to identify top technical talent.

"We can't sit around and just do what we did yesterday to be ready for tomorrow," Barry says. "We've seen enough to know what's possible."